**Transcriptome and genome sequencing uncovers functional variation in human populations**

Tuuli Lappalainen1,2,3#, Michael Sammeth4,5\*, Marc R Friedländer5,6\*, Peter AC ‘t Hoen7\*, Jean Monlong5\*, Manuel A Rivas8\*, Mar Gonzàlez-Porta9, Natalja Kurbatova9, Thasso Griebel4, Pedro G Ferreira5,6, Matthias Barann10, Thomas Wieland11, Liliana Greger9, Maarten van Iterson7, Jonas Almlöf12, Paolo Ribeca4, Irina Pulyakhina7, Daniela Esser10, Thomas Giger1, Andrew Tikhonov9, Marc Sultan13, Gabrielle Bertier5,6, Daniel G MacArthur14,15, Monkol Lek14,15, Esther Lizano5,6, Henk PJ Buermans7,16, Ismael Padioleau1,2,3, Thomas Schwarzmayr11, Olof Karlberg12, Halit Ongen1,2,3, Sergi Beltran4, Marta Gut4, Katja Kahlem4, Vyacheslav Amstislavskiy13, Matti Pirinen8, Stephen B Montgomery1†, Peter Donnelly8, Mark I McCarthy8,17, Paul Flicek9, Tim M Strom11,18, The Geuvadis Consortium, Hans Lehrach13,19, Stefan Schreiber10, Ralf Sudbrak13,19, Ángel Carracedo20, Stylianos E Antonarakis1,2, Robert Häsler10, Ann-Christine Syvänen12, Gert-Jan van Ommen7, Alvis Brazma9, Thomas Meitinger11,18, Philip Rosenstiel10, Roderic Guigó5,6, Ivo G Gut4, Xavier Estivill5,6, Emmanouil T Dermitzakis1,2,3#

\* These authors contributed equally to this work

# Corresponding authors

1 Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland

2 Institute for Genetics and Genomics in Geneva (1G3), University of Geneva, 1211 Geneva, Switzerland

3 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland

4 Centro Nacional d'Anàlisi Genòmica, 08028 Barcelona, Spain

5 Center for Genomic Regulation (CRG, 08003 Barcelona, Spain

6 Pompeu Fabra University (UPF, 08003 Barcelona, Spain

7 Department for Human and Clinical Genetics, Leiden University Medical Center, 2300 RC Leiden, the Netherlands

8 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom

9 European Bioinformatics Institute, EMBL-EBI, Hinxton, United Kingdom

10 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany

11 Institute of Human Genetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany

12 Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, 751 85 Uppsala, Sweden

13 Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

14 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

15 Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge MA 02142, USA

16 Leiden Genome Technology Center, 2300 RC Leiden, the Netherlands

17 Oxford Centre for DIabetes Endocrinology and Metabolism, University of Oxford, Oxford OX3 7BN, United Kingdom

18 Institute of Human Genetics, Technische Universität München, 81675 Munich, Germany

19 Dahlem Centre for Genome Research and Medical Systems Biology, 14195 Berlin, Germany

20 Fundacion Publica Galega de Medicina Xenomica SERGAS, Genomic Medicine Group CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

† Present address: Departments of Pathology and Genetics, Stanford University, 300 Pasteur Dr, Stanford, CA, USA, 94305-5324

**Summary**

Genome sequencing projects are discovering millions of genetic variants in humans. Interpretation of their functional effects is still a challenge, but essential for understanding the genetic basis of variation in human traits. Towards this goal, we sequenced mRNA and small RNA from lymphoblastoid cell lines of 465 individuals from the 1000 Genomes Project. The integration of RNA and DNA sequencing data allowed us to link gene expression and genetic variation, and to characterize mRNA and miRNA variation in several human populations. Common and rare regulatory variants are extremely widespread and affect expression and transcript structure of most genes. By integration of functional annotation we inferred putative causal regulatory variants allowing us to also predict causal variants for dozens of disease-associated loci. Analysis of transcriptome effects of predicted loss-of-function variants uncovered mechanisms for splicing and nonsense-mediated decay. Altogether, this study takes us beyond cataloguing putative functional loci towards understanding and predicting the molecular genomic interactions of causal variants and their cellular effects in the human genome.

**Introduction**

Interpreting functional consequences of genetic variants is one of the biggest challenges in human genomics, as large genome sequencing studies have revealed tens of millions of variants with mostly unknown effects.1 While many studies have successfully linked genetic loci to various human phenotypes, e.g. by genome-wide association studies,2 this usually gives little insight about causal variants and biological mechanisms underlying phenotypic variability and disease susceptibility, despite the improving annotation of the human genome3. One approach to address this challenge has been to analyze cellular phenotypes, such as gene expression, resulting in large catalogs of regulatory variants4-7 known to affect many human diseases and traits.8,9 A major technical advance in these studies has been RNA sequencing, not only allowing characterization of gene expression levels but also uncovering more complex features of transcriptome variation.10,11

The aim of the present study was to characterize transcriptome variation in several human populations by RNA-sequencing of hundreds of individuals, and integrate this with existing high-quality genome sequencing data1 of the same individuals in order to uncover functional variation in the human genome.

**Study design and data set**

In this study, we combined transcriptome and genome sequencing data by performing mRNA and small RNA sequencing on 465 lymphoblastoid cell line (LCL) samples from 5 populations of the 1000 Genomes Project: the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI) (Figure S1, Table S1). Of these samples, 423 were part of the 1000 Genomes Phase 1 dataset1 with genome sequencing data, and the remaining 42 were imputed from Omni 2.5M SNP array data (Supplementary Methods); furthermore we did functional reannotation for all the 1000 Genomes variants (Table S2). Thus, we obtained a powerful dataset to study transcriptome variation and its genetic causes (Fig. S2, S3).

As a parallel goal, we sought to assess the feasibility of distributed RNA sequencing. Thus, transcriptome sequencing was performed in seven European laboratories on randomly allocated RNA samples using the Illumina HiSeq2000 platform, with paired-end 75bp reads for mRNA-seq and single-end 36bp reads for small RNA-seq. We mapped the reads with GEM12 and miraligner13 for mRNA and small RNAs , respectively, resulting in an average of 48.9M well-mapped mRNA-seq reads and 1.2M good-quality micro-RNA (miRNA) reads per sample (Fig. S4; Supplementary Methods). Numerous transcript features were quantified: protein-coding and lincRNA genes (16,084 detected in >50% of samples), transcripts (67,603), exons (146,498), annotated splice junctions (129,805), transcribed repetitive elements (47,409), chimeric transcripts (122 across all samples), previously validated RNA edited sites (99), and mature miRNAs (715) (Table S3, Supplementary methods).

In the final data set after quality control (Supplementary Methods), we had 462 and 452 individuals with good-quality mRNA and miRNA sequencing data, respectively: 91 & 87 (mRNA & miRNA) in CEU, 95 & 93 in FIN, 94 & 94 in GBR, 93 & 89 in TSI, and 89 & 89 in YRI (Table S1). The samples had uniform clustering and distributions of gene and read counts both before and after normalization (Fig. S5-8; ‘t Hoen et al. in preparation). Five samples were prepared (sequencing libraries) and sequenced in replicate in each of the seven laboratories and demonstrated significantly less technical variation among laboratories than biological variation for both mRNA and miRNA data (Mann-Whitney p = < 2.2 × 10-16 for mRNA, p = 1.34 × 10-10 for miRNA; Fig. 1a, S9; ‘t Hoen et al. in preparation). This indicates that, if proper attention is paid to harmonization of protocols and assignment of samples to avoid confounding, RNA sequencing is a mature and robust technology ready for big studies with distributed data production.

In each sample, 11,470-13,300 genes are expressed at >1 RPKM, but since almost every individual turns out to express some unique genes, this adds up to a total of 20,149 out of the 42,958 gene loci in the annotation. Our replicate data shows that that population diversity accounts for 5% of the gene discovery (Fig S10). Thus, large-scale population sequencing uncovers gene expression variation that could not be found in deep sequencing of one or a few samples.

**Transcriptome variation in human populations**

Protein-coding genes differ markedly in how much of the total variation between individuals is explained by overall expression levels and how much by differences in transcript ratios,14 with the latter being more pronounced in most genes (Fig 1b, S11). The proportion of variation explained by expression levels in each gene is highly consistent in different populations (rho=0.84-0.87, Fig S12), which indicates that this pattern of variation is characteristic for each gene and not due to random noise. As expected, the vast majority of the total transcription variation is among individuals within populations, with only 3% explained by population differences. Yet, between population pairs, we detect 263-4379 genes with significant differences in expression levels and/or transcript ratios (Sammeth et al. in preparation). Interestingly, YRI-EUR population pairs have much higher proportions of genes with different transcript usage than European population pairs (75-85% versus 6-40%; Fig. 1c, Fig. S12). Thus, transcript structure variation appears to contribute disproportionally to continental differences, suggesting a special role of splicing in human adaptation.

Notably, not all transcript variation is caused by alternative usage of annotated exons. Within each individual, we observe subtle 1-30 bp variation in a median of 22% of annotated splice donor and acceptor sites with >=10 total reads, with an overrepresentation of in-frame shifts in multiples of 3 bp (Fig S13). This is likely due to common NAGNAG and GYNGYN tandem splice-site patterns as well as noisy splicing15 that leads to loss of transcripts where the reading frame is disrupted through nonsense-mediated decay.16,17 Of major transcript rearrangements we quantified chimeric transcripts with parts from two annotated genes,18 finding 122 of which 78 are seen in all populations (Fig S14, Table S4). Eighty-five percent (85%) of the fusion junctions are less than 100 kb apart, and we found no genetic structural variants corresponding to the junctions with <7.5 kb margin, suggesting that the majority of chimeric transcripts are driven by splicing-related mechanisms rather than genomic rearrangements.

We quantify a total of 644 autosomal miRNAs in >50% individuals, with the 29 highest expressed accounting for 90% of the total miRNA reads (SFig 15). After summing miRNAs with the same seed into 100 families, 36 of them have significant expression level associations with their sets of predicted target exons (P<0.001, global test with Holm’s correction for multiple testing) (Fig 1d; Table S5). Although most mRNAs are targeted by only a single miRNA family, the network is still highly connected. As expected, many miRNA families have mostly negative correlations with mRNA targets. For example, the set of negatively associated targets of the miR-148/152 family is enriched in mRNAs involved in ubiquitination, including *CAND1*, a previously confirmed target for miR148a19. Surprisingly, however, some families like that of miR-181 have mostly positive associations. Interestingly, we find miRNA binding proteins such as Argonaute4 (*EIF2C4*), *BAZ2A*, *BAZ2B* with positive association with miR-181 expression being also negative targets of miR-148/152. Such associations may be a consequence of feed-back mechanisms after the formation of miRNA:protein complexes.

**Transcriptome QTL discovery**

Combining transcriptome and genome sequencing data enables a powerful analysis of regulatory genetic variants. To this end, we mapped cis-QTLs to various transcriptome traits using linear regression separately in the European (EUR) and Yoruba (YRI) populations using genetic variants with >5% frequency in 1MB window and normalized quantifications. Permutations were used to adjust FDR to 5% (Supplementary Methods).

We discovered 7,877 genes with an expression QTL (eQTL) using exon quantifications as the quantitative trait and correcting for the number of exons per gene (out of total 12,981; Table 1), and in the text, we refer to these as eQTLs unless otherwise specified. Regressing out the best associated variant from the EUR eQTL data leaves 7.0% of the exons with another independent eQTL, while as much as 34% of the genes have additional eQTLs for any of their exons, independent from the first association. This indicates substantial genetic independence between exons of the same gene (Fig S16), likely driven by splicing variation that exon eQTLs can capture in addition to gene-wide expression level changes. Indeed, using total gene expression level (RPKM) as the quantitative trait yields only 3773 gene eQTLs out of 13,703 genes. To characterize genetic effects specifically on splicing, we mapped transcript ratio QTLs (trQTLs) using the ratio of each transcript of the total gene quantification, finding a total of 639 genes with a trQTL out of 7,855 analyzed genes – the lower number than for gene eQTLs is not necessarily biological, since total expression level quantification is likely to be more accurate, resulting in higher power. We further analyzed the overlap of gene eQTL and trQTL signals in the EUR sample for the 7,855 genes. Even though these quantitative traits are expected to be biologically orthogonal, we find a significant enrichment of genes with both types of QTL (279 genes = 45% of trQTL genes = 2.15× enrichment, χ2 p < 2.2 × 10-16). Interestingly, regressing out the best trQTL from the gene eQTL analysis showed that in a minimum of 57% of the shared genes, the gene eQTL and trQTL causal variants are not linked (Fig. S17). These analyses imply that while genes may have variants affecting both transcriptional activity and transcript choice, these traits are usually controlled by different genetic elements. The transcript differences driven by EUR trQTLs involve splicing changes only in a minority of cases, with 15% of the transcripts differing in exon skipping and as much as 48% and 43% vary in 5’ and 3’ ends, respectively (note that one trQTLs gene can belong to several categories; Fig 2b). Altogether, these results demonstrate substantial regulatory complexity within individual genes, as well as ubiquitous regulatory variation in the genome with the majority of the genes in our sample – 8384 out of 13970 analyzed genes – having QTLs either for exon or gene expression level or for transcript ratio.

For detection of cis-eQTLs in unannotated transcript elements, we quantified expressed retrotransposon-derived elements outside genes that are known to be an important source for evolution of new exons and transcripts.20 We detected a total of 5933 elements outside annotated genes with significant cis-eQTLs, with an overrepresentation of long terminal repeat retrotransposons (Fig S18). We detected widespread sharing of eQTLs for exons and repeats: 48% and 6% of the best repeat eQTLs variants in EUR were also significant and the best exon eQTLs, respectively (3.7× and 23× enrichment; Fisher exact test p<2.2 × 10-16 for both) compared to set of null variants matched to repeat eQTL frequency and distance from genes. Furthermore, the repeats with eQTLs had significantly higher coexpression with exons and shorter distance to genes than all repeat elements (3kb versus 4kb; Fig S18; KS p<2.2 × 10-16 for both). This suggests that repeat elements can contribute to transcript structure variation through 3’- and 5’-UTR modifications, and that regulatory elements of nearby genes and genetic variants in them often affect repeat element expression too.

Associations to miRNA expression levels (mirQTLs) were found for 60 out of the 644 miRNAs (Fig S15), showing that genetic effects on miRNA expression are common. While we have limited power for detecting trans-eQTLs, mirQTLs show small signs of negative trans-effects to their targets (Fig. S19). Furthermore, of the 99 bona dfide RNA edited sites reported in the DARNED database21 and quantified in our data (all A-to-I changes, see Supplementary Methods), we found 8 having genetic associations to the degree of editing, suggesting that this trait can also be under genetic control (Fig S20). The editQTLs were located close to the edited site (median 304 bp), which is expected given that editing can be affected by local sequence context22.

**Characterization of regulatory variants**

Genome sequencing data gives us a unique opportunity to characterize causal regulatory variants and their properties. In these analyses, we focused on our large data set of exon eQTLs from Europeans and describe results from these unless otherwise mentioned; results from Yoruba and from trQTLs are shown in Supplementary Figures. To characterize the properties of eQTLs and estimate our ability to find causal variants, we compared the most significant eQTL variant per gene – the most likely causal variant – to the 2nd, 5th and 10th best variants, and to a null distribution of non-eQTL variants matched for distance from transcription start site and minor allele frequency (Figure 2a; Supplementary Methods).

In 13% of the genes the strongest eQTL variant is an indel, which is 37% more than for the matched null variants (Fisher exact test p = 5.6 × 10-6; Fig S21), suggesting that indels are more likely to have functional effects than SNPs. eQTLs are highly enriched in many noncoding elements from the Ensembl Regulatory Build, such as many transcription factor peaks (median enrichment 3.5×, median p = 0.011), histone marks (median 2.7×, median p = 5.56 × 10-9), DNase1 hypersensitive sites (3.7×, p = 1.54 × 10-21), and a high enrichment in chromatin states of active promoters (3.8x, p = 1.28 × 10-35) and strong enhancers (median 2.7×, median p = 4.89 × 10-6) (Fig S22). Within genes, best eQTL variants were most enriched in 5’UTRs (3.9x, p = 4.49 × 10-16) consistently with the neighboring promoter activity, while splice-site (2.3×, p = 5.76 × 10-3) and nonsynonymous (2.4×, p = 2.72 × 10-5) enrichments point to coding variants having also regulatory functions. Transcript ratio QTLs are overrepresented in splice sites (6.8×, p = 2.44 × 10-7), as expected, but also for example in 3’UTRs (2.5×, p = 1.83 × 10-6) and H3K9ac sites (2.9×, p = 6.57 × 10-6) and promoters (2.4×, p = 5.79 × 10-6) (Fig. S23). Altogether, these analyses add to our previous understanding23,24 of the functional consequences of genetic variants to transcription.

These analyses show a characteristic pattern of highest enrichment for the best eQTL variant, and a rapidly decreasing trend towards lower ranks. With this information, we were able to quantify how likely the first variant is to be the causal variant: we calculated the annotation enrichment of the best eQTL variants relative to the matched null for (1) all eQTL loci, and (2) those loci where we reasoned that the best eQTL variant is the causal due to having a log 10 p-value >1.5 higher than the second variant; this threshold was based on saturation of the functional enrichment of the best (1st) versus the 2nd eQTL variant (Fig S24; Supplementary Methods). From the ratio of the enrichments (1) and (2), we estimate the best variant is the causal variant in 57% of EUR eQTLs and 75% of YRI eQTLs. The higher percentage in YRI is expected due to lower LD confounding. Not relying on SNP array data25 is an important advantage in the characterization of causal variants: 15% of the eQTL genes do not have a single Omni 2.5M array SNP among the significant variants, and in 81% the best (i.e. most likely causal) variant is not on the Omni 2.5M chip (Fig 2c, Fig S25).

Our eQTL variants have significant enrichment of GWAS26 variants (16% of 6473 GWAS variants are eQTLs in EUR or YRI, versus 11% of a frequency-matched GWAS null distribution; chi2 p < 2.2 × 10-16; for trQTLs the percentages are 1.8% and 0.84%, respectively, p = 7.2 × 10-9). The large proportion of overlap even under the null warrants caution against inference of causality from a mere overlap of QTL-GWAS signals; however, the significant enrichment of GWAS SNPs in the top eQTL ranks (p=1.97 × 10-6; Fig. S26) suggests that eQTL effects indeed have increased probability to be causal GWAS mechanisms as observed before8,9. We can now apply our ability to predict causal eQTL variants to find causal disease-associating variants for these loci. To this end, we selected 86 strictly filtered eQTL regions that have been statistically estimated to underlie GWAS signals using the RTC method,7,8 and of these, the best eQTL variant in our data – i.e. the most likely causal variant – is not the GWAS SNP in all but one case, and not in previously commonly used genotyping arrays (Affymetrix 500K and Illumina 550) in all but 6 cases (Fig S27). Together, this gives us 80 GWAS associations where we can pinpoint novel putative causal variants, listed in Table S6. Figure 2d shows an example of the DGKD gene locus where an intronic SNP rs838705 is associated to calcium levels27, and 21 kb downstream we find a top eQTL variant – an insertion of two nucleotides – located close the start site of two transcripts and in the middle of regulatory regions for H3K27ac, MEF2A, MEF2C, and DNase1.

**Allele-specific transcription**

RNA sequencing enables analysis of potential regulatory differences between the two haplotypes of an individual. We analyzed allele-specific expression (ASE) – imbalance between the two heterozygous alleles – as well as allele-specific transcript structure (ASTS), which is a *de novo* approach that captures differences in the exonic distribution of reads from the two haplotypes (Fig S2, S28, Supplementary Methods). In our dataset, we analyzed ASE and ASTS in a median of 8420 and 2135 sites per individual, respectively, of which 6.5% and 5.6% have a significant signal (p<0.005) of allelic difference. The substantial overlap of ASE and ASTS signals (Fig 3a) indicates that ASE is often driven by transcript structure variation rather than different overall expression levels of the two haplotypes.

Allelic difference is usually a signal of cis-regulatory variation,10,28 which is supported by clear population structure seen in ASE data (Fig S29, S30). ASE data can thus be used as a proxy especially for rare regulatory variation that cannot be seen in eQTL analysis.28 Indeed, the majority of allele-specific events are rare in the population, and further enriched for high-impact effects (Fig 3b, Fig S31) – suggesting that rare regulatory variants are likely to be an important class of functional variation on humans. However, not all allelic imbalance is driven by noncoding regulatory variants: the likelihood of a site to have ASE or ASTS is dramatically increased (up to 60%) for variant classes related to splicing, probably due to their direct causal effect on transcript structure (Fig 3c).

**Interpretation of loss-of-function variation**

While QTL analyses aim at identifying previously unknown regulatory variants, we can also observe and quantify functional effects of Loss-of-Function (LoF) variants.29 Our samples have 2,987 premature stop codon variants and 4,090 splice-site variants of which 839 and 849, respectively, are captured by our RNA-seq data. Most of these variants are rare and seen in only one or two individuals. First, we analyzed premature stop variants that may induce nonsense-mediated decay.30 As expected, we observe a high enrichment for ASE with loss of the variant allele (Fig 3c, 4a, S32) as in previous studies28,29. Variants close to the end of the transcript are predicted to escape NMD30, which is generally the case in our data. However, of the variants predicted to cause NMD, only 32% (54% of rare variants MAF<1%) are ASE outliers, while in 68% (46%) the ASE results do not support NMD (Fig 4b,c). This suggests currently unknown mechanisms of NMD escape.

To quantify the effect of splicing variants, we calculated a score of the predicted change in splicing efficiency for all genetic variants overlapping an annotated splicing motif (Sammeth et al. in preparation). On average, nonreference variants have lower splicing affinity (p<2.2 × 10-16, Fig. S33), and 10% are predicted to destroy the motif. This effect can be validated by our RNA sequencing data: individuals carrying two predicted splice-motif-destroying alleles have 29% lower median inclusion rates of the affected exon (p<2.2 × 10-16, Fig 4d), indicating that the variants indeed lead to exon skipping.

**Conclusions**

By integrated analysis of RNA and DNA sequencing data we were able to obtain a unique high-precision view to variation of the transcriptome and its genetic causes. Transcriptome sequencing of multiple populations allowed us to estimate the relative contribution of total expression level versus transcript structure variation within and between populations. We detected transcriptome QTLs and allelic transcription affecting expression levels or transcript structure of a large number of genes, and dissected the relationship of different types of regulatory variants in high depth. Furthermore, we demonstrated how transcriptome analysis of mostly rare loss-of-function variants can improve our understanding of this class of variants with high functional relevance. For the first time, we were able to predict large numbers of causal regulatory variants, which is a key to real understanding of the source of functional regulatory variation of the genome. This provides important biological insights to the molecular sources of phenotypic variability, and brings us closer to developing methods for functional interpretation of personal genomes. Ultimately, this study illustrates the power of combining genome sequence analysis with a high-depth functional readout such the transcriptome.

**Methods summary**

Total RNA was extracted from EVB transformed lymphoblastoid cell line pellets by the TRIzol reagent (Ambion), and mRNA and small RNA sequencing of 465 unique individuals was performed on the Illumina HiSeq2000 platform in seven European laboratories, each processing 48-113 randomly assigned samples. Five samples were sequenced in replicate in each of the labs for both mRNA and miRNA, and twice in UNIGE for mRNA. The mRNA reads were mapped with GEM12 to hg19, Gencode v1231 exons were quantified from read overlaps, and FluxCapacitor10 was used for transcript quantification. Small RNA data was mapped and quantified with miraligner13 using miRBase v1832. Quantifications were normalized by the total number of well-mapped reads. Data quality was assessed by sample correlations and read and gene count distributions, and technical variation was removed by PEER normalization33 for the QTL and miRNA-mRNA correlation analyses.

**References**

1 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:nature11632 [pii]

10.1038/nature11632 (2012).

2 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:nature05911 [pii]

10.1038/nature05911 (2007).

3 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:nature11247 [pii]

10.1038/nature11247 (2012).

4 Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-428, doi:nature06758 [pii]

10.1038/nature06758 (2008).

5 Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-1224, doi:ng2142 [pii]

10.1038/ng2142 (2007).

6 Liang, D. M. & Qiao, J. J. [Recent advances in the study of macrolide glycosyltransferases]. *Yao Xue Xue Bao* **42**, 455-462 (2007).

7 Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-1089, doi:ng.2394 [pii]

10.1038/ng.2394 (2012).

8 Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895, doi:10.1371/journal.pgen.1000895 (2010).

9 Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888, doi:10.1371/journal.pgen.1000888 (2010).

10 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:nature08903 [pii]

10.1038/nature08903 (2010).

11 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772, doi:nature08872 [pii]

10.1038/nature08872 (2010).

12 Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, doi:nmeth.2221 [pii]

10.1038/nmeth.2221 (2012).

13 Pantano, L., Estivill, X. & Marti, E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res* **38**, e34, doi:gkp1127 [pii]

10.1093/nar/gkp1127 (2010).

14 Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res* **22**, 528-538, doi:gr.121947.111 [pii]

10.1101/gr.121947.111 (2012).

15 Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**, e1001236, doi:10.1371/journal.pgen.1001236 (2010).

16 Hiller, M. *et al.* Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet* **36**, 1255-1257, doi:ng1469 [pii]

10.1038/ng1469 (2004).

17 Carter, M. S., Li, S. & Wilkinson, M. F. A splicing-dependent regulatory mechanism that detects translation signals. *EMBO J* **15**, 5965-5975 (1996).

18 Parra, G. *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* **16**, 37-44, doi:gr.4145906 [pii]

10.1101/gr.4145906 (2006).

19 Murata, T. *et al.* miR-148a is an androgen-responsive microRNA that promotes LNCaP prostate cell growth by repressing its target CAND1 expression. *Prostate Cancer Prostatic Dis* **13**, 356-361, doi:pcan201032 [pii]

10.1038/pcan.2010.32 (2010).

20 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703, doi:nrg2640 [pii]

10.1038/nrg2640 (2009).

21 Kiran, A. & Baranov, P. V. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**, 1772-1776, doi:btq285 [pii]

10.1093/bioinformatics/btq285 (2010).

22 Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**, 321-349, doi:10.1146/annurev-biochem-060208-105251 (2010).

23 Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-394, doi:nature10808 [pii]

10.1038/nature10808 (2012).

24 Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol* **13**, R7, doi:gb-2012-13-1-r7 [pii]

10.1186/gb-2012-13-1-r7 (2012).

25 Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* **4**, e1000214, doi:10.1371/journal.pgen.1000214 (2008).

26 Hindorff, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies (2010).

27 O'Seaghdha, C. M. *et al.* Common variants in the calcium-sensing receptor gene are associated with total serum calcium levels. *Hum Mol Genet* **19**, 4296-4303, doi:ddq342 [pii]

10.1093/hmg/ddq342 (2010).

28 Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **7**, e1002144, doi:10.1371/journal.pgen.1002144

PGENETICS-D-10-00589 [pii] (2011).

29 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:335/6070/823 [pii]

10.1126/science.1215040 (2012).

30 Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* **23**, 198-199, doi:S0968-0004(98)01208-0 [pii] (1998).

31 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:22/9/1760 [pii]

10.1101/gr.135350.111 (2012).

32 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157, doi:gkq1027 [pii]

10.1093/nar/gkq1027 (2011).

33 Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770, doi:10.1371/journal.pcbi.1000770 (2010).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Author Contributions**

Designed the study: TL, TGi, SBM, PACH, EL, HL, SS, RS, AC, SEA, RH, ACS, GJvO, AB, TM, Pro, RG, IGG, XE, ETD

Coordinated the project: TL, TGi, GB, XE, ETD

Participated in data production: TL, TGi, IPa, MSu, EL, SB, MG, VA, KK, DE, PR, OK

Analyzed the data: TL, MSa, MRF, PACH, JM, MAR, MGo, NK, TGr, PGF, MB, TW, LG, MvI, JA, PRi, IPu, DE, AT, MSu, DGM, ML, EL, HPJB, IPa, TS, OK, HO, SB, MGu, KK, VA, MP, PD, MIM, PF, TMS

Drafted the paper: TL, ETD

**Data access**

All RNA sequencing data of the project as well as genotypes with our functional re-annotation of all the 1000 Genomes Phase 1 variants  are available via ArrayExpress and European Nucleotide Archive (ArrayExpress accession numbers E-GEUV-1, E-GEUV-2, E-GEUV-3). The project wiki is open to the public in geuvadiswiki.crg.es. For the visualisation of RNA-sequencing analysis we created the Geuvadis Data Browser ([www.ebi.ac.uk/Tools/geuvadis-das](http://www.ebi.ac.uk/Tools/geuvadis-das)) where exon and miRNA quantifications and eQTLs can be viewed and searched (Fig. S34).

Reprints and permissions information is available at www.nature.com/reprints

Correspondence and requests for materials should be addressed to tuuli.e.lappalainen@gmail.com or emmanouil.dermitzakis@unige.ch

**Acknowledgements**

We would like to thank Emilie Falconnet, Luciana Romano, Alexandra Planchon, Deborah Bielsen, Alisa Yurovsky, Alfonso Buil, Julies Bryois, Alexandra Nica, Johan Rung, Nikolay Kolesnikov, Asier Roa, Eugene Bragin, Simon Brent, Justo Gonzalez, Marta Morell, Anna Puig, Emilio Palumbo, Marina Ventayol Garcia, Jeroen F.J. Laros, Julie Blanc, Rahnehild Birkelund, Gloria Plaja, Matt Ingham, Jordi Camps, Monica Bayes, Lidia Agueda, Anais Gouin, Marie-Laure Yaspo, Elisabeth Graf, Anett Walther, Carola Fischer, Sandy Loesecke, Bianca Schmick, Daniela Balzereit, Simon Dökel, Matthias Linser, Alexander Kovacsovics,  Melanie Friskovec, Catharina von der Lancken, Melanie Schlapkohl, Anita Dietsch, Markus Schilhabel, the SNP&SEQ Technology Platform in Uppsala, Sascha Sauer (Max Plank Institute for Human Genetics (MPIMG, Berlin, Germany) for ESGI (European Sequencing and Genotyping Infrastructure) coordination, the Vital-IT high-performance computing center of the SIB Swiss Institute of Bioinformatics, Bernadette Goldstein and others at the Coriell Institute, and James Cooper, Edward Burnett, Karen Ball and others at the European Collection of Cell Cultures (ECACC) and the 1000 Genomes Consortium.

This project was funded by the European Commission 7th Framework Program, Project N. 261123 (GEUVADIS), as well as The Swiss National Science Foundation 31003A\_130342 and 127375, the Louis Jeantet Foundation, NIH-NIMH MH090941, ERC 249968 grant, The Swedish Research Council (90559401), the Knut and Alice Wallenberg Foundation (2011.0073), the DFG Cluster of Excellence Inflammation at Interfaces, the INTERREG4A project HIT-ID, the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project, grant agreement HEALTH-F4-2007-201413, the Centre for Medical Systems Biology within the framework of The Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO), Spanish Plan Nacional funding SAF2008-­00357 (NOVADIS), the Generalitat de Catalunya funding AGAUR 2009 SGR-1502, the Ministry of Economy and Competitiveness funding to CLL-Consortium project, ESGI – European Commission Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 262055, Academy of Finland, Emil Aaltonen Foundation, EMBO long-term fellowship ALTF 225-2011, European Community’s FP7 75 CAGEKID (grant agreement 241669), FIS PS09/02368, and the German Ministry of Education and Research (01GR0802, 01GM0867, 01GR0804).

**Tables**

**Table 1**: Quantitative trait locus discovery. Numbers of transcriptome features with a QTL (FDR 5%). Exons and transcripts are collapsed to the gene level. Union refers to nonredundant count in EUR and YRI.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Total** | **EUR (n=373)** | **YRI (n=89)** | **Union** |
| **exon eQTL** | 12981 genes | 7486 | 2308 | 7877 |
| **gene eQTL** | 13703 genes | 3259 | 501 | 3773 |
| **transcript ratio QTL** | 7855 genes | 620 | 83 | 639 |
| **mirQTL** | 644 miRNAs | 57 | 15 | 60 |
| **repeat eQTL** | 43875 repeats | 5652 | 993 | 5933 |
| **RNA edit QTL** | 99 edited sites | 8 | 0 | 8 |

**Figure Legends**

**Figure 1: Transcriptome variation.** a) Spearman rank correlation of replicate samples, based on mRNA exon and miRNA quantifications of 5 individuals sequenced 8 and 7 times for mRNA and miRNA, respectively. The boxplots are separated by the individual (INDV) and the sequencing lab being the same or different (DIFF). The quantifications have been normalized only for the total number of mapped reads (see Fig. S9 for correlations after normalization). b) The proportion of expression level variation (as opposed to splicing) of the total transcription variation between individuals in each population, measured per gene. c) Proportion of genes with significant differential expression levels and/or transcript usage between population pairs. The numbers on the right denote the total number of differential genes. d) Network of 35 significant (P<0.001) miRNA families and their significantly associated mRNA targets (P<0.05). The edges display negative (green) and positive (orange) associations.

**Figure 2. Transcriptome QTLs.** Functional annotation of EUR eQTLs (a), with an enrichment of eQTLs in regulatory and coding annotations for the 1st, 2nd, 5th and 10th best associating eQTL variant per gene, relative to a matched null set of variants. The vertical line at 1 denotes the null, and the numbers are –log10 p-values of a Fisher test between the best eQTL and the null set for each category. b) The rank of the best Omni2.5M SNP among the significant EUR eQTL variants per gene, in bins on the x-axis according to the total number of significant variants. c) Classification of changes caused by transcript ratio QTLs. d) eQTL associations in the EUR population for the STAT6 gene, with the SNP rs167769 associating to eosinophilic esophagitis shown in red and the top eQTL variant and the likely causal disease variant rs4559 shown in blue.

**Figure 3: Allele-specific effects on expression and transcript structure.** a) Sharing of allele-specific expression (ASE) and transcript structure (ASTS) signals: For significant (p<0.005) ASE sites, the distribution of ASTS p-value of the sites in the same individual is plotted, and vice versa for significant ASTS sites. The ASE p-values are calculated from sites sampled to exactly 30 reads in order to avoid inflation of sharing due to shared high coverage. The numbers denote the pi1 statistic measuring the enrichment of low p-values. b) Frequency of significant ASE event in the population (x-axis) and the effect size (abs(0.5 – REF/TOTAL)) of the significant ASE events, calculated per ASE SNP. Only ASE SNPs with >=20 heterozygote individuals with >=30 reads were included, and the data was corrected for coverage biases and false positives by sampling and permutations (see Supplementary Methods). d) Frequency of ASE and ASTS by annotation class shows a clear enrichment of allelic imbalance in loss-of-function sites.   
  
  
**Figure 4: Transcriptome effects of loss-of-function variants.** Nonsense-mediated decay due to premature stop codon variants was measured using allele-specific expression. a) shows the distribution of non-reference allele ratios (on the y-axis) for premature stop variants, and the scatterplots break this distribution to individual variants sorted on the x-axis according to derived allele frequency, with sites classified to those predicted to trigger (b) and escape (c) NMD. The dots denote the median across individuals, and the vertical lines show the range of ratios for variants carried by several individuals. The grey vertical lines denote derived allele frequencies of 0, 0.001 and 0.01. d) Exon inclusion scores for variable exons for individuals that carry 0, 1 or 2 copies of variants that destroy a splice motif.